Enriching the Human Phenotype Ontology with inferred axioms from textual descriptions*

Shahad Kudama[1], Rafael Berlanga[1], Ernesto Jiménez-Ruiz[2]

[1] Jaume I University, Castellón, Spain
[2] University of Oslo, Oslo, Norway

Abstract. The Human Phenotype Ontology (HP) is a reference vocabulary of human phenotypic abnormalities. HP, apart from the textual information (general definitions, descriptions, synonyms, etc.) of each ontology concept, also provides computer-readable logical definitions (axioms) of terms that will allow human phenotypic abnormalities to be related to entities from anatomy, pathology, biochemistry and other areas. In this paper we present a prototype to generate new axiomatic knowledge from the textual descriptions of each HP term. The prototype *(i)* detects terms in the textual descriptions and not found in the given logical expressions, *(ii)* generates pair combinations of those terms, *(iii)* builds triples after detecting the most probable relation between the pair of terms using a statistical model and, finally, *(iv)* suggests the most probable triples to the user so she can decide which ones can be added to the original axioms.

1 Introduction

The large amount of public knowledge resources available in the Web have been developed regardless of the processing and integration needs of modern information systems and, this fact, is obstructing its massive use. We clearly need richer lexicons and axiomatic knowledge resources.

In this paper, we focus on the axiomatic knowledge resources and address the problem of how to exploit these resources to improve processing and integration tasks. The research is done in the technological context of the Semantic Web, because one of its main objectives is to generate semantic annotations from knowledge resources. In particular, there is special interest in the information processing in the phenotype and genotype field, where descriptions tend to be logical representations that allow inferencing over them. The main challenge is how to exploit semantic properties of these resources in the processing and analysis.

In this paper, we rely on the Human Phenotype Ontology (HP) [1] and the annotation facilities provided by BioPortal [2].

HP is an ontology, expressed in the Web Ontology Language (OWL, a family of knowledge representation languages for authoring ontologies) [3], that aims at providing a standardized vocabulary of phenotypic abnormalities related to human disease. Each term in the HP describes a phenotypic abnormality, such as *'atrial septal defect'*. It currently contains approximately $11,000$ terms and over $115,000$ annotations to hereditary diseases. It also includes axioms for the terms, which are a formal way

* This work was partially funded by the BIGMED project (IKT 259055), the SIRIUS Centre for Scalable Data Access (Research Council of Norway, project no.: 237889).

to describe taxonomies and classification networks, essentially defining the structure of knowledge for various domains: the nouns representing classes of objects and the verbs representing relations between the objects. We focus on the task of extracting axioms, from textual descriptions of phenotypes. We summarize the main objectives of the work presented in this paper as follows:

- Analysis of the HP axioms to understand how relations between HP classes are expressed.
- Use of the descriptive textual annotations of the HP classes and detect terms that are not being used in the related axioms.
- Design of a statistical model to infer relations between a given pair of ontology classes.
- Use of the statistical model to generate a list of triples (subject, relation, object) ranked based on the probability of having this relation between the two concepts.
- Select the most probable triples and propose this new knowledge for a subsequent (manual) assessment to convert valid and relevant triples into suitable HP axioms.

The processing of free text and the discovery of implicit relations arise as two of the most important challenges in this paper. On the one hand, free text brings ambiguity and vagueness. On the other hand, potential relations between classes will most probably not be explicitly expressed as verbs in the textual information, and thus the relation names will need to be inferred.

2 Related work

In this section, we briefly introduce some approaches relevant to the work presented in this paper. In [4], an effort to elucidate Obol (Open Bio-Ontology Language) is carried out and the attempts to reason over the resulting definitions are presented. [5, 6] represent efforts to normalize the Gene Ontology [7] in a way that can be better exploited by reasoners. Thanks to the logical definitions of an ontology, we can gradually begin to automate many aspects of ontology development, detecting errors and filling in missing relationships. Another related work is the Semantic Medline, more specifically the SemMedDB [8]. This approach aims at building a triple store of semantic annotations in UMLS that are extracted from predications identified in PubMed abstracts. These predications are associated to the predefined set of relationships of the UMLS Semantic Network. Unfortunately, the tool for extracting predications (i.e., SemRep) depends on specific versions of UMLS Metathesaurus, which are not freely available and do not have the same domain coverage as BioPortal. Moreover, the relations provided by SemRep are different from those used in HP and BioPortal, as discussed later in Section 4.

3 Methodology

We have represented the axioms in HP as triples (subject, relation, object), in order to generate a statistical model and being able to infer the most suitable relation to each (subject, object) pair of annotations. Apart from the axioms, the HP ontology contains, for each class or term, a set of lexical metadata: definition, description, synonyms, etc. We extracted and annotated the lexica using the BioPortal annotator,[3] an online service

[3] `https://bioportal.bioontology.org/annotator`

that discover annotations for biomedical texts with classes from different ontologies stored in BioPortal [2]. Using the extracted annotations, we generated for each HP class a list of pairs (subject, object) with all possible combinations of the annotations.

We give, as input, to the statistical model the list of pairs (subject, object) and we obtain the same list enriched with relations (subject, relation, object), ranked by the probability given by the statistical model. The last step is guided by the user, he is the person able to review the list of ranked triples, decide which ones are useful and, finally, build new axioms to be added to the HP classes.

3.1 Transforming axioms into triples

We extracted from HP all textual information and all the axioms for each term. Then we expressed axioms in an easier way to make use of them, by implementing a method for parsing and transforming OWL axioms defining a class to observation triples (statistic units): subject, relation, object. Here we can see an example of the different stages to go from a Description Logic axiom to a set of triples. Starting with, for example, the following axiom (belonging to *HP_0000871*, *Panhypopituitarism*):

```
['SomeValuesFrom(BFO_0000051
     IntersectionOf(PATO_0000462
        SomeValuesFrom(RO_0000052
            IntersectionOf(
                 IntersectionOf(GO_0003008
                     SomeValuesFrom(BFO_0000066 UBERON_0002196))
                 SomeValuesFrom(BFO_0000050 UBERON_0000468)))
        SomeValuesFrom(RO_0002573 PATO_0000460)))']
```

We focused only the classes involved in the axiom (the concrete OWL constructor or restriction is not relevant for our approach), and we worked just with the name of the ontology, not the term code. For example, (*HP_0100752*, *UBERON_3010224*) is changed to (*HP*, *UBERON*). The reduction of axioms and the corresponding triples are shown in the table below.

['BFO'	(HP_0000871, BFO, PATO)
['PATO',	(PATO, RO, GO)
['RO',	(GO, BFO, UBERON)
[['GO', ['BFO', 'UBERON']],	(RO, BFO, UBERON)
['BFO', 'UBERON']],	(PATO, RO, BFO)
['RO', 'PATO']]]	(PATO, RO, PATO)

3.2 Generating the statistical models for axioms

After moving from each axiom to a set of triples, abstracting the ontological information, we estimate the probabilities between the different components of these triples. More specifically, our aim is to estimate the following marginal distributions: $P(s^*|r)$ for subject-relation pairs, $P(o^*|r)$ for object-relation pairs, and $P(r|s^*, o^*)$ for relation against subject-object pairs. When estimating these probabilities, we abstract s and o to their component ontologies (denoted with $.^*$ superscript) so that we can rank relation schemas. With the previous distributions, we can rank the inferred triples for each pair extracted from the textual descriptions. We use the maximum likelihood estimation (MLE), using factorization as follows:

$$P(s^*, r, o^*) = P(r|s^*) \cdot P(r|o^*) \cdot P(r|s^*, o^*)$$

3.3 Generating new knowledge through semantic annotation

We annotated the HP descriptions of each concept by using BioPortal. With these annotations, all possible triples are generated by combining pairs of annotations that co-occur in each sentence of the description and all the potential relations that can hold between them. We also add constraints over the entities to be related. For example, both subject and object have to be in the same sentence and subject or object (or both of them) is not in the given axioms, so we can be sure that new triples are adding knowledge. Finally, by using the statistical model, candidate triples for each HP are ranked.

4 Results

Results are provided as triples (subject, relation, object), representing knowledge that is not present within the axioms associated to the HP classes. We obtained $76,348$ new triples for $8,582$ HP classes, so the number of triples we infer for each HP class is, on average, 8.[4]

As an example, we have the term *HP_0100752* with preferred label *'Hepatic anomalous lobulation'*. Currently, this term does not have any (direct) axiom associated in the HP. The term *HP_0100752* also has the following textual descriptions: *'Anomalous liver lobulation'*, *'Abnormal liver lobulation'* and *'Formation of abnormal lobules (small masses of tissue) in the liver'*. After using the proposed method, we obtained the following triples:

subject	*relation*	*object*	*prob.*
masses	inheres in	liver	0.18081
masses	inheres in	tissue	0.18081
abnormal	inheres in	liver	0.18081
abnormal	inheres in	tissue	0.18081
masses	has modifier	abnormal	0.12623
tissue	part of	liver	0.02376

We have performed a preliminary evaluation by comparing the extracted triples against the SemMedDB predictions [8].[5] For this purpose, we crossed extracted triples and predications by subject and object. As a result, we were able to match 41,200 (54%) triples to SemMedDB predications. This indicates that our approach generates meaningful triples. We also inspected non-matched triples, and many of them can be considered correct. However, a strict evaluation must be performed to assess their true accuracy. As for relations, SemMedDB deals with a much richer set of relations compared to HP. Analyzing the matched triples-predications, the main identified alignments between SemMedDB and HP relations are: (*PART_OF*, *part of*), (*OCCURS_IN*, *inheres in*), (*ASSOCIATED_WITH*, *inheres in*) and (*AFFECTS*, *has modifier*). However, SemMedDB and HP relationships are not easily comparable as they are used in different ways. This issue deserves an in-deep study in the future work.

[4] Raw results: `http://krono.act.uji.es/swat4ls_2017/results.txt`

[5] SemMedDB predictions: `http://krono.act.uji.es/swat4ls_2017/evaluation.txt`

To sum up, results are promising as we are able to extract knowledge that it is not explicitly present in the axioms. With this knowledge experts should decide which extracted triples are useful for her, and then, create and add new axioms associated to the HP classes.

5 Conclusions

Many efforts have been done to give structure and formal definitions to biomedical ontologies, which enable the use of reasoners in order to infer (implicit) knowledge from the ontology. There is still, however, plenty of work to do in this area as the domain keep evolving and the ontologies need to keep track of this new knowledge in a coherent and complete manner. The maximum potential of any ontology will be obtained when all its terms have a complete and exhaustive set of logical definitions.

In this paper, we have presented a method to enrich the logical information available in the HP classes. Using a statistical model and extracting the missing concepts in the axioms, the system proposes a list of candidate triples that can be used by experts to build new axioms. We have compared the generated triples with the SemMedDB predications, showing a notable overlap between them and therefore their meaningfulness.

As future work, we plan to define further relevance criteria for providing a better ranking of triples, as only using probability thresholds do not give us always good results. We also need to design and implement a solid and complete evaluation process as the task of doing it manually is not manageable, due to the large amount of data we are dealing with. We also plan to make use of the alignment between HP and UMLS to obtain a richer lexicon associated to HP classes, because BioPortal annotations are often too short and consequently, they do not cover the full semantics of the text. Finally, the system could build the axioms automatically and be able to tune the statistical model with the user feedback, considering that accepting or rejecting a triple is a valuable information to be used as input of the statistical model.

References

1. Kohler, S., et al.: The Human Phenotype Ontology in 2017. Nucleic Acids Research **45** (2017) D865–D876
2. Noy, N.F., et al.: Bioportal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Research **37**(Web-Server-Issue) (2009) 170–173
3. Consortium, W.W.W.: OWL 2 Web Ontology Language document overview. W3C (2009)
4. Mungall, C.J.: Obol: Integrating language and meaning in bio-ontologies. Willey InterScience (2004) 509–520
5. Mungall, C.J., et al.: Cross-product extensions of the gene ontology. Journal of Biomedical Informatics **44** (2011) 80–86
6. Wroe, C., Stevens, R., Goble, C.A., Ashburner, M.: A methodology to migrate the gene ontology to a description logic environment using DAML+OIL. In: PSB. (2003)
7. Ashburner, M., et al.: Creating the gene ontology resource. design and implementation. Willey InterScience (2001) 425–433
8. Kilicoglu, H., et al.: SemMedDB: a PubMed-scale repository of biomedical semantic predications. Bioinformatics **28**(23) (2012)